Are Any-to-Any Models More Consistent Across Modality Transfers Than Specialists?

Jiwan Chung **Janghan Yoon** **Junhyeong Park** **Sangeyl Lee**
Joowon Yang **Sooyeon Park** **Youngjae Yu**
Yonsei University

jiwan.chung.research@gmail.com

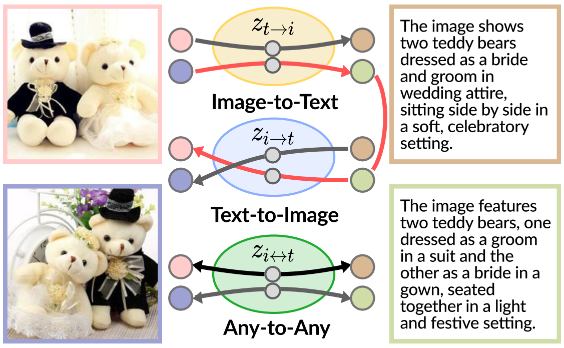
Abstract

Any-to-any generative models aim to enable seamless interpretation and generation across multiple modalities within a unified framework, yet their ability to preserve relationships across modalities remains uncertain. Do unified models truly achieve cross-modal coherence, or is this coherence merely perceived? To explore this, we introduce ACON, a dataset of 1,000 images (500 newly contributed) paired with captions, editing instructions, and Q&A pairs to evaluate cross-modal transfers rigorously. Using three consistency criteria—cyclic consistency, forward equivariance, and conjugated equivariance—our experiments reveal that any-to-any models do not consistently demonstrate greater cross-modal consistency than specialized models in pointwise evaluations such as cyclic consistency. However, equivariance evaluations uncover weak but observable consistency through structured analyses of the intermediate latent space enabled by multiple editing operations. We release our code and data at https://github.com/JiwanChung/ACON.

1 Introduction

Any-to-any generative models are designed to both interpret and generate multiple modalities—such as text, images, and audio—within a unified framework (Wang et al., 2022; Lu et al., 2024; Wang et al., 2024). In contrast to modality-specific approaches which often rely on textual interfaces to mediate generation (OpenAI, 2023; Lu, 2024), any-to-any models share the majority of parameters across different modalities. This design choice suggests potential advantages in flexibility and transferability between modalities.

However, the practical value of any-to-any models remains uncertain. At their current stage of development, they often fail to consistently outperform specialized models (Podell et al., 2024; Labs, 2024; Liu et al., 2024) in terms of output quality. Also, they may be less training-efficient due to the significant computational overhead of optimizing a single, large-scale system. As a result, it remains unclear whether such models confer tangible benefits over their modality-specific counterparts.



The image shows two teddy bears dressed as a bride and groom in wedding attire, sitting side by side in a soft, celebratory setting.

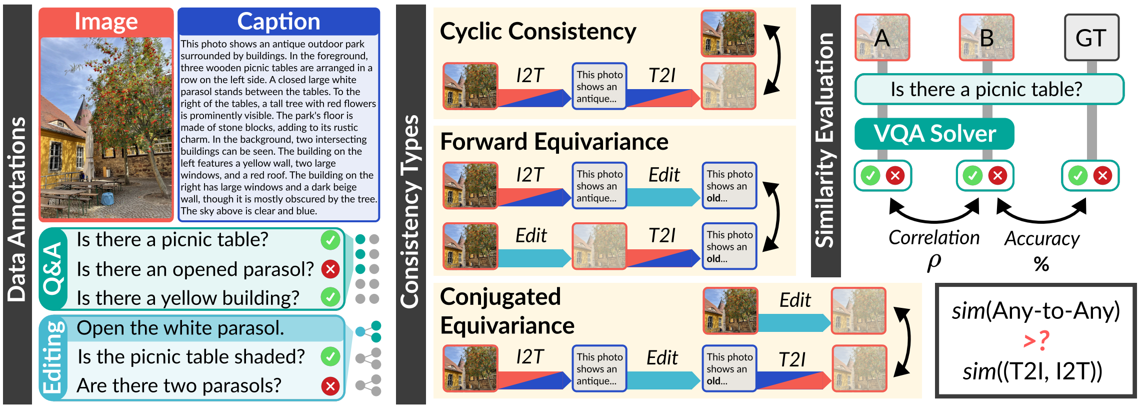
The image features two teddy bears, one dressed as a groom in a suit and the other as a bride in a gown, seated together in a light and festive setting.

Figure 1: We examine the consistency of any-to-any models compared to separate image-to-text and text-to-image models. An effective any-to-any model, capable of learning a unified latent space z, is expected to mitigate issues like cyclic consistency failures, as depicted by the red lines. The illustration is a conceptual case drawn with images from MMVP (Tong et al., 2024).

What, then, should we anticipate from any-to-any models? Prior work (Huang et al., 2021; Lu, 2023) has proposed viewing multimodal learning as an attempt to approximate a shared latent representation from each modality's partial view. Building on this perspective, we posit that if a single any-to-any model successfully learns such a unified latent space, it should produce more coherent cross-modal conversions than two separate modality-specific models, each constrained by its own distinct latent approximation.

To verify this conjecture, we test whether any-to-any models achieve greater consistency in cross-modal transfer than pairs of modality-specific models. We formalize this consistency using three criteria: *cyclic consistency*, requiring that converting an input from text to image and back again recovers the original input; *forward equivariance*,

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2595–2606
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



This photo shows an antique outdoor park surrounded by buildings. In the foreground, three wooden picnic tables are arranged in a row on the left side. A closed large white parasol stands between the tables. To the right of the tables, a tall tree with red flowers is prominently visible. The park's floor is made of stone blocks, adding to its rustic charm. In the background, two intersecting buildings can be seen. The building on the left features a yellow wall, two large windows, and a red roof. The building on the right has large windows and a dark beige wall, though it is mostly obscured by the tree. The sky above is clear and blue.

Figure 2: We evaluate whether any-to-any modality conversion models demonstrate greater consistency compared to unrelated pairs of independent image-to-text and text-to-image generators. (Left) To this end, we curate a dataset with detailed annotations, including captions, Q&As, and editing prompts. (Centre) Consistency is measured using three criteria: cyclic consistency, forward equivariance, and conjugated equivariance, where the latter two require off-the-shelf image or text editing tools for in-modality transformations. (Right) The evaluation involves comparing the similarity between two generated outputs (images or textual descriptions) using an external VQA solver to compute correlation (ρ) and accuracy (%).

ensuring that applying a modification before or after cross-modal conversion yields the same result; and *conjugated equivariance*, providing an alternative formulation of equivariance that utilizes both text-to-image and image-to-text conversions.

We introduce ACON (Any-to-any CONsistency), a meticulously annotated dataset to assess coherence in cross-modal transformations. It comprises 1,000 images, including 500 private images specifically contributed for this study. Each image is paired with a human-written dense caption aimed at faithful reconstruction, three image-editing instructions, and ten binary question-answer pairs for evaluating output similarity. In addition, every editing instruction is accompanied by two prompt-conditioned Q&As to capture the effects of the transformation.

Experiments on ACON reveal that any-to-any models do not consistently exhibit greater cross-modal consistency compared to arbitrary combinations of specialized models, particularly in pointwise evaluations such as cyclic consistency. However, equivariance evaluations demonstrate that weak consistency between text-to-image and image-to-text capabilities can be observed in distributional analyses of the intermediate latent space, enabled by multiple editing operations.

We anticipate that ACON will serve as a diagnostic benchmark to evaluate the benefits and trade-offs of training any-to-any models within a unified framework: any-to-any models do not show great consistency at the current stage. We will release our code and data to support further research and development in this field.

2 Defining Consistency Across Modalities

We formalize the concept of multimodal consistency for numerical experiments by adopting three widely recognized types of consistency: cyclic consistency, equivariance to transformation, and commutativity of operations.

Notations Let ϕ denote the parameters of a text-to-image generation model, and ψ denote the parameters of an image-to-text model. We define two types of operations:

1. Across-Modality *Conversion* ($f(x)$): Transformations between modalities, such as generating an image from text ($f^{t\to i}$) or generating text from an image ($f^{i\to t}$).

2. In-Modality *Modification* ($g(x,p)$): Edits or modifications within the same modality, such as image editing (g^{i}) with a given prompt p. To implement g, we use off-the-shelf LLMs (g^{t}) and image editing models (g^{i}) because existing any-to-any models are not optimized for editing, and our focus is on evaluating cross-modality consistency rather than in-modality performance.

Furthermore, a data sample x consists of two views: an image view (x^{i}) and a text view (x^{t}). If a single model is used for both modality directions, it follows that $\phi = \psi$.

Cyclic Consistency is a commonly used concept in machine learning, particularly in unpaired translation tasks (Zhu et al., 2017; Bielawski and VanRullen, 2023). It ensures that applying transformations between two domains consecutively returns the input to its original state. For example, in unpaired image-to-image translation, translating an image from domain A to domain B and back to domain A should reconstruct the original image.

In our setup, cyclic consistency ensures that transformations between text and image modalities are invertible. Specifically:

$$f_\psi^{i\to t}(f_\phi^{t\to i}(x^t)) = x^t \tag{1}$$

$$f_\phi^{t\to i}(f_\psi^{i\to t}(x^i)) = x^i \tag{2}$$

Forward Equivariance is a property that ensures the consistency of transformations under secondary operations (Cohen and Welling, 2016). Specifically, applying a modification to the input followed by a transformation should yield the same result as applying the transformation first, followed by the corresponding modification.

In this work, we adapt the traditional definition by treating modification in both modalities equivalently ($g^i \simeq g^t$). In our context, this principle ensures compatibility between in-modality modifications and across-modality conversions:

$$f_\phi^{t\to i}(g^t(x^t, p)) = g^i(f_\phi^{t\to i}(x^t), p), \tag{3}$$

$$f_\psi^{i\to t}(g^i(x^i, p)) = g^t(f_\psi^{i\to t}(x^i), p). \tag{4}$$

Note that forward equivariance compares transformations within the same direction, such as $f_\phi^{t\to i}$ applied before or after a modification. We thus incorporate another form of equivariance relation which uses both directions at a time to evaluate consistency between modality conversions in the following paragraph.

Conjugated Equivariance extends the idea of forward equivariance by incorporating transformations in both directions to evaluate consistency across modality conversions. Specifically, we modify forward equivariance by inverting f in the left terms, which yields:

$$f_\psi^{i\to t}(g^i(f_\phi^{t\to i}(x^t), p)) = g^t(x^t, p) \tag{5}$$

$$f_\phi^{t\to i}(g^t(f_\psi^{i\to t}(x^i), p)) = g^i(x^i, p) \tag{6}$$

Conjugated equivariance can also be seen as an extension of cyclic consistency, where the intermediate latent space is explicitly modified before completing the transformation cycle. By incorporating multiple modifications, this approach extends pointwise evaluations to analyze structural multi-point consistency within the shared latent space.

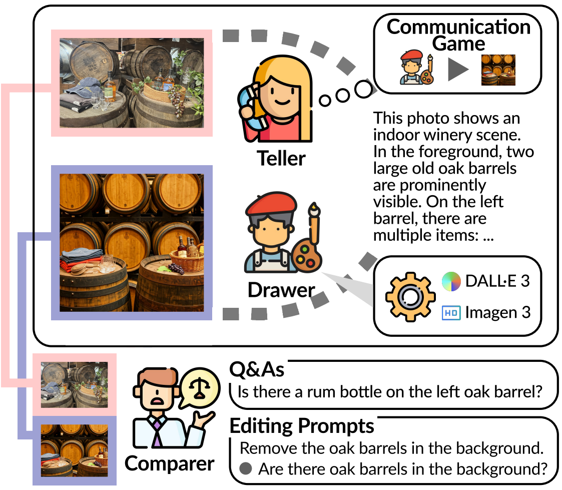

Communication Game
Teller
Drawer
This photo shows an indoor winery scene. In the foreground, two large old oak barrels are prominently visible. On the left barrel, there are multiple items: ...
DALL·E 3
Imagen 3
Comparer
Q&As
Is there a rum bottle on the left oak barrel?
Editing Prompts
Remove the oak barrels in the background.
Are there oak barrels in the background?

Figure 3: Data annotation process for ACON. Three human workers perform distinct roles: the *teller* creates a textual description emphasizing key elements for reconstruction, following a communication game framework (Kim et al., 2019), the *drawer* recreates the image using the description via multi-turn AI interactions, and the *comparer* generates Q&As to capture similarities and differences between the original and reconstructed images, annotating editing instructions as well.

3 Data Collection Process

To support the operations outlined in section 2, we curated a dataset comprising 1,000 image inputs (i), corresponding text captions (t), 3,000 editing prompts (p), and 16,000 question-answer pairs designed to evaluate similarity between images or captions. Further details on the data collection methodology, including human resourcing, can be found in appendix B.

Images The input images i are curated from both *unseen* and *seen* sources to ensure diversity and relevance. For the *unseen* subset, we collect 500 images that have not been exposed to any available MLLMs during training. Volunteers from the research community contributed private photos, which were manually filtered based on the following criteria: 1) exclusion of images with potential privacy violations or toxic content; 2) removal of images requiring domain-specific skills, such as named entity recognition or OCR capabilities; 3) elimination of low-quality images, such as those with motion blur, small resolution, or skewed aspect ratios; and 4) exclusion of overly simplistic content with very few objects. This filtering process resulted in the removal of approximately 76% of the original submissions. For the *seen* subset, we randomly sampled 500 images from the widely used COCO Captions dataset (Chen et al., 2015), which provides a benchmark set of images commonly utilized in multimodal learning.

Captions The goal of our captioning process is to create captions that accurately guide the reconstruction of the original image. We manually annotate dense captions t for input images i, employing the communication game framework (Kim et al., 2019) to ensure alignment between modalities. Annotators are assigned three roles: the *teller*, the *drawer*, and the *judge*.

The teller crafts a caption t that encodes the essential visual details of the image i, acting as the sender in the communication game. The drawer interprets the caption and reconstructs the image using tool-assisted generation (DALL·E-3 (Betker et al., 2023) and Imagen 2 (Google DeepMind, 2023)), functioning as the receiver. The reconstructed image $\hat{i}$ is compared to the original image i to evaluate the success of the communication. To maintain objectivity, the teller does not have access to the reconstructed image during annotation, ensuring that captions are crafted independently of the reconstruction process. Finally, the *judge* evaluates the quality of the reconstructed image and provides feedback (good/bad). If necessary, re-captioning is requested from a different *teller*.

Questions & Answers A reliable metric is essential to assess factual similarity between images. Retrieval-based metrics, such as CLIPScore (Hessel et al., 2021), are insufficient for evaluating factual correctness, particularly in compositionality (Ma et al., 2023) or counting tasks (Radford et al., 2021). Recent studies (Hu et al., 2023; Cho et al., 2024) propose an alternative approach: generating questions that capture salient facts about the images. One such metric, VQAScore (Lin et al., 2024), automatically generates questions using a pretrained VQA question generator and then scores consistency by comparing model answers on reference and generated images. While VQAScore improves over retrieval-based methods by grounding evaluation in factual content, its reliability is limited by the quality and scope of automatically generated questions, which tend to be shallow and generic, failing to stress fine-grained or compositional differences.

By contrast, our approach introduces a human-in-the-loop *comparer*, who carefully constructs challenging questions that highlight both similarities and differences between the reference and reconstructed images. These questions are specifically designed to distinguish subtle failures in object placement, count, or relational semantics that automatic systems often overlook. For each image pair, five similarity- and five difference-oriented questions are created, ensuring coverage of both aligned and misaligned aspects. This deliberate design leads to more sensitive and discriminative factual evaluations than automatic pipelines such as VQAScore.

Editing Operations To evaluate equivariance and commutativity properties, we define modification operations for each modality (g^i for images and g^t for text). These operations are conditioned on a prompt p, specifying the nature and direction of the modification. The *comparer* annotates three prompts per image, ensuring alignment with the intended changes, and additionally generates two prompt-conditioned question-answer pairs per editing prompt to reflect the specific modifications.

Manual Filtering To address quality variance inherent in collaborative annotation, we conducted a rigorous manual filtering process aimed at normalizing differences across annotators. Observed inconsistencies included variation in the verbosity of image descriptions, factual correctness, and formatting conventions in Q&As (e.g., inconsistent use of parentheses to denote objects or attributes). To ensure consistency, we established strict filtering criteria: (1) questions must be answerable based solely on the provided description, (2) answers must be factually correct, (3) the question set must be diverse and cover different object types or visual properties, and (4) descriptions must contain sufficient detail. Any factual inconsistencies led to automatic rejection.

Before initiating the filtering process, all reviewers (distinct from the original annotators) participated in a calibration phase to align evaluation standards. Each reviewer shared and critiqued 10 annotated examples with others, enabling discussion on interpretation and enforcement of the filtering criteria. After normalization, each sample was independently reviewed by two human judges. Approximately 43% of the initial samples were discarded and replaced with new annotations that met all quality standards.

4 Experiments

4.1 Setups

This section outlines the models tested and the utilities employed in our experiments. Full details, including model checkpoints and instruction prompts, are available in appendix A and appendix D.

Models For *text-to-image* generation, we use the open-source models including Flux (Labs, 2024) and Stable Diffusion XL (Podell et al., 2024). These models were selected for their balance between performance and resource efficiency, allowing the evaluation to focus on consistency rather than absolute image fidelity. For *image-to-text* generation, we include LLaVA-Next (Liu et al., 2024) (abbreviated as LLaVA) and Qwen2VL (Bai et al., 2023). These models are chosen for their widespread use, robust performance across tasks, and practical trade-offs in computational requirements. To assess the central claim that *any-to-any models* improve cross-modal coherence, we evaluate four open-source systems: Chameleon (Team, 2024), Emu-3 (Wang et al., 2024), VILA-U (Wu et al., 2024), and Seed-X (Ge et al., 2024b). For model descriptions, refer to section 5.

Utilities *In-Modality Editors*: For editing within a single modality, we use Cos Stable Diffusion XL 1.0 Edit (CosXL) (Podell et al., 2024) for image editing and Qwen2.5 (Yang et al., 2024) for text editing. *Evaluators*: For visual question-answering tasks, we employ PaliGemma2 (Steiner et al., 2024) for its strong performance in static VQA scenarios. We use Qwen2.5 for textual Q&As.

4.2 Cyclic Consistency

Cyclic consistency refers to the model's capability to accurately reconstruct input data by leveraging its latent representations, ensuring the preservation of original content through a bidirectional transformation process. For image reconstruction, the process involves an image-to-text transformation followed by a text-to-image transformation ($x \xrightarrow{f^{i\to t}} z \xrightarrow{f^{t\to i}} \bar{x}$). For text reconstruction, the order is reversed ($x \xrightarrow{f^{t\to i}} z \xrightarrow{f^{i\to t}} \bar{x}$). The reconstructed data $\bar{x}$ is compared to the original input x for evaluation.

Metrics We employ off-the-shelf visual or textual question-answering tools to compare the reconstructed data $\bar{x}$ with the original input $x \in \mathcal{X}$. Given the model output $\bar{x}$, the context $c \in \mathcal{C}$, and a question $q \in \mathcal{Q}$, evaluation is conducted using a parameterized binary classifier $h_\theta : \mathcal{X} \times \mathcal{C} \times \mathcal{Q} \to \{0, 1\}$. The primary accuracy metric is defined as:

$$sim(x, \bar{x}) := \sum_q \delta\big(h_\theta(\bar{x}, c, q), h_o(x, c, q)\big), \quad (7)$$

where δ is the Dirac delta function, and h_o represents the oracle classifier that provides the ground-truth labels. Note that we average over ten different questions per model-generated output. In addition to accuracy, we report the F1-score, which incorporates precision and recall, to assess the similarity between the binary outputs.

Results The cyclic consistency evaluation results are presented in table 1. Notably, a single any-to-any model does not consistently outperform the combination of separate specialized models in cyclic consistency, raising questions about the presumed advantages of training a single any-to-any model.

Notable exceptions include Seed-X and VILA-U, which demonstrate notable consistency when utilizing a single any-to-any model. In contrast, other any-to-any models such as Chameleon and Emu3 fail to exhibit consistent patterns. This disparity aligns with the visual tokenization strategies employed by these models: both Seed-X and VILA-U adopt *semantically-aligned* visual tokenizers, either by leveraging features from a pre-trained ViT or by optimizing alignment with textual representations. On the other hand, Chameleon and Emu3 rely solely on image reconstruction objectives. This finding indicates that incorporating semantic modeling into visual tokenization may contribute to improved alignment of the latent space during modality conversions.

Still, the results show that this evaluation conflates per-modality transfer performance with cyclic consistency. For instance, VILA-U, when used as the secondary text-to-image operator, achieves high performance regardless of the initial operation it is paired with. A similar trend is observed in text generation, where models such as LLaVA and Qwen2VL tend to outperform others. In conclusion, evaluating any-to-any consistency requires multiple complementary criteria, which we address in the following experiments.

4.3 Forward Equivariance

Forward Equivariance assesses the impact of applying an in-modality editing operation (g) either

	Image → Text → Image				Text → Image → Text		
I2T	**T2I**	**Accuracy (%)**	**F1 (%)**	**T2I**	**I2T**	**Accuracy (%)**	**F1 (%)**
LLaVA	Flux	61.78	72.29	Flux	LLaVA	63.73	70.28
	SDXL	57.21	67.91		Qwen2VL	**66.93**	**73.61**
	Chameleon	56.13	66.77		Chameleon	56.42	61.95
	Emu3	60.91	71.65		Emu3	63.48	70.03
	Seed-X	**62.57**	**73.37**		Seed-X	63.52	70.22
	VILA-U	61.40	71.99		VILA-U	62.36	68.59
Qwen2VL	Flux	62.86	73.20	SDXL	LLaVA	59.20	65.10
	SDXL	57.83	68.53		Qwen2VL	58.04	63.98
	Chameleon	57.04	67.65		Chameleon	57.17	62.63
	Emu3	61.50	73.28		Emu3	57.58	63.07
	Seed-X	62.65	73.33		Seed-X	57.29	62.67
	VILA-U	**63.36**	**73.92**		VILA-U	**59.70**	**65.43**
Chameleon	Flux	55.38	65.75	Chameleon	LLaVA	55.63	**60.82**
	SDXL	54.47	64.85		Qwen2VL	54.38	59.06
	Chameleon	53.93	64.33		Chameleon	**55.76**	60.59
	Emu3	55.81	66.23		Emu3	53.97	58.66
	Seed-X	57.04	68.05		Seed-X	54.59	59.32
	VILA-U	**59.04**	**69.79**		VILA-U	55.46	60.24
Emu3	Flux	58.00	68.63	Emu3	LLaVA	**62.61**	**68.75**
	SDXL	54.01	64.28		Qwen2VL	61.40	67.75
	Chameleon	53.72	63.85		Chameleon	57.46	63.06
	Emu3	58.62	69.20		Emu3	60.03	66.17
	Seed-X	59.04	69.79		Seed-X	61.15	67.23
	VILA-U	**59.29**	**69.98**		VILA-U	59.20	64.93
Seed-X	Flux	60.91	71.39	Seed-X	LLaVA	59.95	65.79
	SDXL	57.46	67.94		Qwen2VL	60.95	67.32
	Chameleon	56.13	66.54		Chameleon	56.42	61.76
	Emu3	60.53	71.28		Emu3	57.46	63.09
	Seed-X	**61.78**	**72.53**		Seed-X	**61.45**	**67.61**
	VILA-U	61.07	71.65		VILA-U	59.41	65.12
VILA-U	Flux	60.41	70.97	VILA-U	LLaVA	55.84	60.99
	SDXL	58.37	69.00		Qwen2VL	57.87	63.47
	Chameleon	55.13	65.52		Chameleon	56.25	61.53
	Emu3	60.82	71.66		Emu3	55.46	60.50
	Seed-X	60.70	71.64		Seed-X	56.50	61.94
	VILA-U	**62.15**	**72.88**		VILA-U	**58.12**	**63.69**

Table 1: Cyclic consistency evaluation results. The best scores for the same initial operation are highlighted in **bold**, while the second-best scores are underlined. Results obtained using a single any-to-any model, instead of distinct model pairs, are presented in color.

before or after the modality transfer (f). Unlike other consistency criteria, this approach focuses on comparing outputs from the same modality transfer direction ($f^{i \to t}$ vs. $f^{i \to t}$ and $f^{t \to i}$ vs. $f^{t \to i}$).

Metrics This evaluation involves three key comparisons: the two model outputs ($f(g(x))$ and $g(f(x))$) and the ground-truth modified datapoint x'. The primary metric is the Pearson correlation between $f(g(x))$ and $g(f(x))$, emphasizing consistency over absolute performance. Additional metrics, $sim(f(g(x)), x')$ and $sim(g(f(x)), x')$, are detailed in the appendix.

Similarity between datapoints is measured using question-answering methods, as in the cyclic consistency evaluation. However, each question and answer here is conditioned on an editing prompt $p \in \mathcal{P}$. For each sample, we compute averages over two prompt-conditioned questions per editing prompt, using three editing prompts per image or text.

Results We illustrate correlation statistics in fig. 4, while the complete results are presented in appendix C. The findings reaffirm earlier observations: the consistency of any-to-any models relative to independent specialist pairs is not consistently superior. However, notable exceptions include Seed-X and VILA-U, which exhibit improved textual consistency, consistent with trends observed in previous experiments.

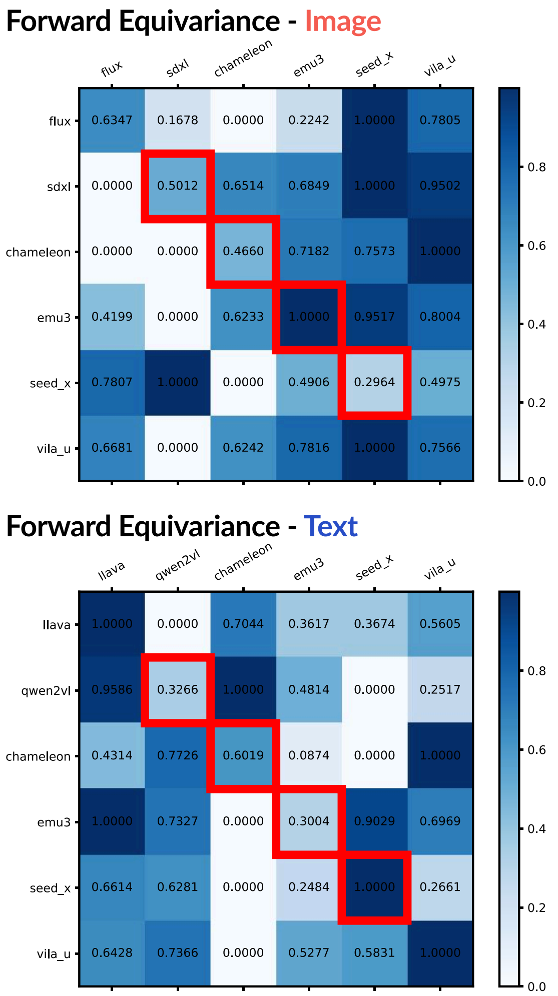


Figure 4: *Correlation* of forward equivariance across model pairs, normalized to [0, 1] per row. Diagonal components with red borders indicate the same any-to-any generator used for both image-to-text and text-to-image transfers.

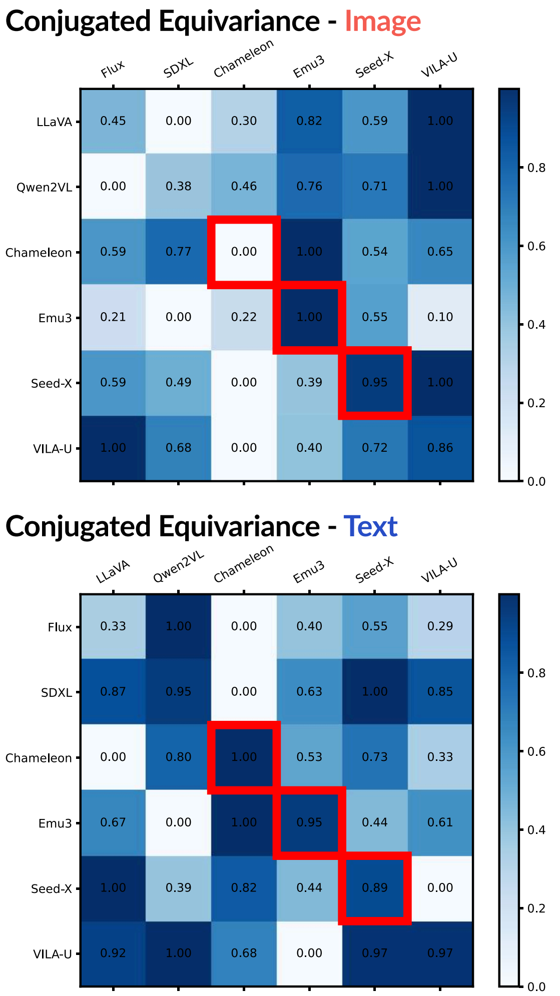


Figure 5: *Accuracy* of conjugated equivariance across model pairs, normalized to [0, 1] per row. Diagonal components with red borders indicate the same any-to-any generator used for both image-to-text and text-to-image transfers.

4.4 Conjugated Equivariance

Conjugated Equivariance extends cyclic consistency by applying an in-modality operation g between modality transfers. Taking image reconstruction as an example, the goal is to reconstruct the image with the modification in the latent textual description represented correctly ($x \xrightarrow{f^{i\to t}} z \xrightarrow{g^t} z' \xrightarrow{f^{t\to i}} \bar{x}'$). This approach shifts the focus from evaluating single-point reconstructions to assessing the alignment of transformation directions (vectors) across modalities. The process is applied analogously for textual reconstruction.

Metrics This evaluation compares two terms: the model output $\bar{x}'$ and the ground-truth label x'. Thus, we report accuracy and F1 score as in the cyclic consistency experiment. The only difference is that the questions are also conditioned on the editing prompts. Thus, we average results over two questions per editing prompt, testing three editing prompts per sample. We do not generate the in-modality output ($g(x, p)$). Instead, we use ground-truth answers to the question to replace the evaluation results ($h_o(g(x, p), c, p)$).

Results Empirical results, visualized in fig. 5 and detailed further in appendix C, reveal consistent self-alignment for most any-to-any models. All any-to-any models, except for Chameleon in image generation, achieve stable self-consistency when paired with themselves. However, these models do not consistently outperform when paired with themselves compared to being paired with other models. This suggests that while any-to-any models demonstrate stable self-alignment, they are not always the optimal choice for their own outputs in cross-modal operations.

This raises the question: why is any-to-any consistency, barely noticeable in cyclic consistency, more evident in the conjugated equivariance experiment? A plausible explanation is that conjugated equivariance evaluates consistency by analyzing transformations across a distribution of edited latent representations, rather than focusing on a single transformation. By leveraging multiple editing operations, this approach captures broader patterns of alignment in the latent space, enabling a more nuanced assessment of consistency between text-to-image and image-to-text capabilities. This finding aligns with the shared latent learning hypothesis (Huang et al., 2021; Lu, 2023), which posits that models trained on multiple tasks or modalities form a unified latent space for shared representations.

4.5 Qualitative Results

Figure 6 presents sample inference results used to evaluate cyclic consistency and conjugated equivariance. A key observation is the stylistic disparity between natural (ground truth) and model-generated images, underscoring the limitations of cyclic consistency as a reliability metric. This supports the use of equivariance, which directly compares outputs across models and avoids distortions arising from differences between natural and synthetic images. Furthermore, Chameleon's poor image fidelity often aligns with object composition errors (e.g., incorrect counts or misplaced elements), highlighting that any-to-any models do not consistently translate their stronger linguistic capabilities into accurate compositional image generation.

4.6 Discussion

Diversity This work does not explicitly address the diversity of generated images or text, as we employ deterministic sampling throughout. While distributional analysis would be ideal for evaluating coverage of semantic space, its effectiveness is constrained by the limited generative diversity of current image synthesis models. As noted in prior work (Hsieh et al., 2024), models such as Stable Diffusion XL (Podell et al., 2024) often fail to produce semantically distinct outputs even when conditioned on different random seeds, limiting the effectiveness of stochastic sampling for diversity evaluation.

5 Related Work

Any-to-Any Models Any-to-any generative models aim to unify multimodal understanding and generation across diverse tasks and modalities. Here, we focus on image and text modalities to align with the scope of this paper. Recent approaches can be categorized into deterministic and distributional modeling of image data. Deterministic approaches, such as Kosmos-G (Pan et al., 2024) and Emu2 (Sun et al., 2024a), directly regress CLIP (Radford et al., 2021) features, which are then fed into an (optionally fine-tuned) Stable Diffusion (Rombach et al., 2022) generator. Distributional approaches, by contrast, often compress images into discrete token sequences using vector quantization (van den Oord et al., 2017), enabling categorical latent space modeling. Examples include OFA (Wang et al., 2022), Unified-IO 2 (Lu et al., 2024), Chameleon (Team, 2024), LaVIT (Jin et al., 2023), and Emu3 (Wang et al., 2024). To enhance information sharing between text-to-image and image-to-text tasks, recent models such as SEED-LLaMA (Ge et al., 2024a), SEED-X (Ge et al., 2024b), and VILA-U (Wu et al., 2024) incorporate semantic alignment into their tokenization strategies. Additionally, diffusion-based approaches, exemplified by Transfusion (Zhou et al., 2024), are emerging as alternatives to categorical tokenization, leveraging continuous distributions for greater flexibility.

Multimodal Consistency Multimodal consistency ensures coherence across modalities. MM-R3 (Chou et al., 2024) and MMCBench (Zhang et al., 2024a) evaluate robustness to semantic shifts and corrupted inputs. Advances in text-to-image consistency include PDF-GAN (Tan et al., 2022), which employs Semantic Similarity Distance, and a diffusion framework (Sun et al., 2024b) leveraging knowledge graphs. MC-MKE (Zhang et al., 2024b) addresses modality errors, while ConsiStory (Tewel et al., 2024) improves layout consistency without additional training. CycleGAN (Bielawski and VanRullen, 2023) and CyclePrompt (Diesendruck et al., 2024) enhance captioning and code generation with cycle-supervised methods. Semantic consistency metrics (Bent, 2024) and cycle-consistency losses (Zhu et al., 2017) further refine reliability.

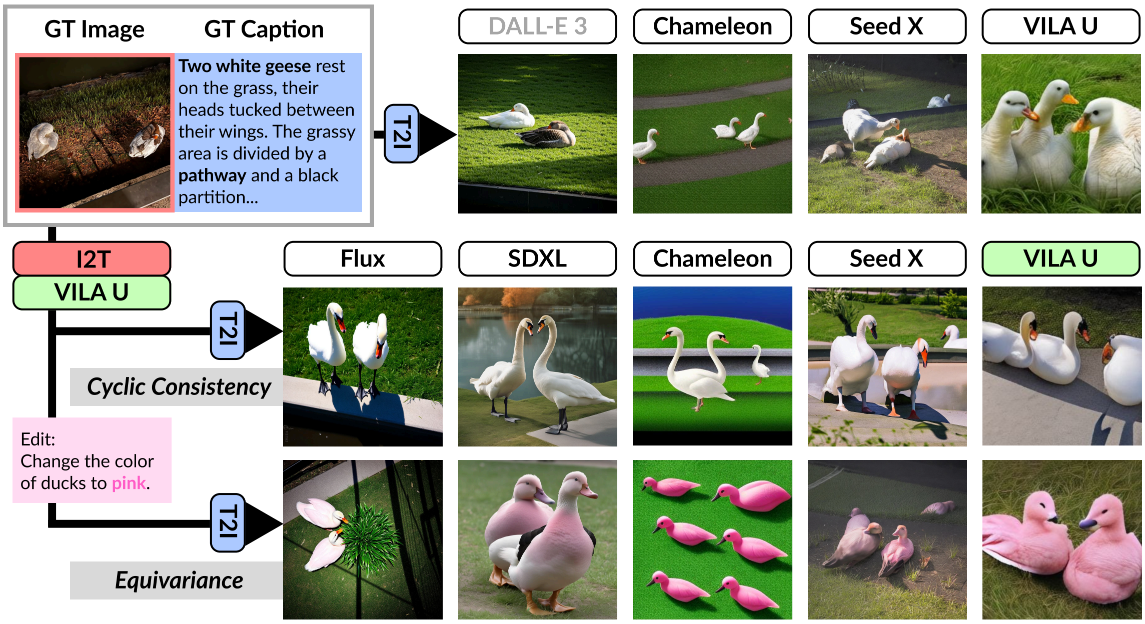

GT Image
GT Caption
Two white geese rest on the grass, their heads tucked between their wings. The grassy area is divided by a pathway and a black partition...
T2I
DALL-E 3
Chameleon
Seed X
VILA U
I2T
VILA U
T2I
Flux
SDXL
Chameleon
Seed X
VILA U
Cyclic Consistency
Edit:
Change the color of ducks to pink.
T2I
Equivariance

Figure 6: Example inference results using VILA-U (Wu et al., 2024) as an example image-to-text captioner. (Top row) Text-to-image transfer results are shown using the human-annotated caption as input for reference. (Middle row) *Cyclic consistency* in the image domain is evaluated by captioning the original image with the captioner, then reconstructing it using different image generators. (Bottom row) *Conjugated Equivariance* is assessed by applying an editing operation g in the latent domain before reconstructing the image.

6 Conclusion

We introduce ACON, a hand-annotated benchmark designed to evaluate the any-to-any consistency of multimodal AI models. Our analysis reveals that existing any-to-any models exhibit weak consistency between text-to-image and image-to-text tasks, which becomes apparent only through distributional inspections of the intermediate latent space, facilitated by multiple editing operations.

7 Limitations & Future Directions

Limitations Our experiments are conducted using any-to-any models in their *as-is* state. Since model behavior results from the interplay of data, architecture, and training processes, this monolithic evaluation does not allow us to isolate the specific factors contributing to (in)consistency across modalities. We encourage the research community, particularly those with greater computational resources, to undertake controlled analyses to systematically examine how each design component of any-to-any models impacts their consistency.

Our new benchmark, ACON, has certain limitations stemming from its curation process, which focuses on hand-taken natural images. This emphasis impacts the dataset's image distribution in several ways. First, our private images exclude artistic images, 2D drawings, and 3D renderings, limiting the scope for evaluating any-to-any consistency in these domains. Second, as the dataset relies on pre-taken image contributions, the subjects are predominantly confined to realistic scenarios typically captured by people, such as scenic landscapes, food, or animals. Although efforts were made to ensure diversity, these inherent distributional biases persist in the dataset.

Additionally, ACON was annotated by five NLP researchers sharing similar cultural backgrounds. Although a separate group of human evaluators validated these annotations, we acknowledge the potential influence of cultural bias on image descriptions. For instance, studies (Nisbett et al., 2001; Ananthram et al., 2024) suggest that individuals from Western cultures often emphasize the central figure in an image, while those from Eastern cultures are more inclined to consider the broader scene context.

Future Directions Future directions include:

1. **Iterative Composition**: This work focuses on a single cyclic loop of modality transfers. Exploring iterative composition of transfers could provide further insights into consistency. Neural networks approximating data distribu-

tions are known to collapse output diversity under repeated application—would consistent cyclic loops mitigate this effect?

2. **Extending Modalities**: Expanding beyond the (image, text) modality pair to others, such as (speech, text), could uncover whether any-to-any models demonstrate stronger consistency across different domains.

Risks We introduce a new multimodal data corpus, including newly contributed private images. Each image has undergone manual inspection to prevent copyright infringement, portrait rights violations, and the inclusion of harmful or inappropriate content. However, some risks remain:

- **Bias and Representational Gaps**: Despite efforts to ensure diversity, the dataset may inadvertently overrepresent or underrepresent certain cultural or demographic backgrounds, potentially leading to biased model outputs or unfair generalizations.
- **Unintended Personal Data Exposure**: While we obtained explicit consent from contributors and filtered out any images that could reveal their identity, advancements in AI, such as geographic inference models, may enable the extraction of private information from images in unintended and non-explicit ways.
- **Erosion of Zero-Shot Integrity**: By releasing new private images, we aim to encourage evaluation on truly unseen data. However, public availability of the dataset risks its use for fine-tuning future models, potentially compromising the integrity of results in subsequent zero-shot evaluations.

8 Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2024-00457882, AI Research Hub Project; No.RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); and No.RS-2025-02263598, Development of Self-Evolving Embodied AGI Platform Technology through Real-World Experience) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00354218).

References

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Brinnae Bent. 2024. Semantic approach to quantifying the consistency of diffusion model image generation. *arXiv preprint arXiv:2404.08799*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Romain Bielawski and Rufin VanRullen. 2023. Clip-based image captioning via unsupervised cycle-consistency in the latent space. In *8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 266–275. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*.

Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *The Twelfth International Conference on Learning Representations*.

Shih-Han Chou, Shivam Chandhok, James J Little, and Leonid Sigal. 2024. MM-R^3: On (in-) consistency of multi-modal large language models (mllms). *arXiv preprint arXiv:2410.04778*.

Taco Cohen and Max Welling. 2016. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR.

Maurice Diesendruck, Jianzhe Lin, Shima Imani, Gayathri Mahalingam, Mingyang Xu, and Jie Zhao. 2024. Learning how to ask: Cycle-consistency refines prompts in multimodal foundation models. *arXiv preprint arXiv:2402.08756*.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024a. Making LLaMA SEE and draw with SEED tokenizer. In *The Twelfth International Conference on Learning Representations*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024b. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Google DeepMind. 2023. Imagen 2.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Yu-Guan Hsieh, Cheng-Yu Hsieh, Shih-Ying Yeh, Louis Béthune, Hadi Pour Ansari, Pavan Kumar Anasosalu Vasu, Chun-Liang Li, Ranjay Krishna, Oncel Tuzel, and Marco Cuturi. 2024. Graph-based captioning: Enhancing visual descriptions by interconnecting region captions. *arXiv preprint arXiv:2407.06723*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. 2023. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Black Forest Labs. 2024. Flux.1: Advanced text-to-image generation. `https://blackforestlabs.ai/flux-1/`.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455.

Pengqi Lu. 2024. Qwen2vl-flux: Unifying image and text guidance for controllable image generation.

Zhou Lu. 2023. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36:57244–57255.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Richard E Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. 2001. Culture and systems of thought: holistic versus analytic cognition. *Psychological review*, 108(2):291.

OpenAI. 2023. Gpt-4: Openai's large multimodal model. `https://openai.com/research/gpt-4`.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024a. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Yichen Sun, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024b. Prompt-consistency image generation (pcig): A unified framework integrating llms, knowledge graphs, and controllable diffusion models. *arXiv preprint arXiv:2406.16333*.

Zhaorui Tan, Xi Yang, Zihan Ye, Qiufeng Wang, Yuyao Yan, Anh Nguyen, and Kaizhu Huang. 2022. Ssd: Towards better text-image consistency metric in text-to-image generation. *arXiv preprint arXiv:2210.15235*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. 2024. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024a. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943*.

Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, and Xiaojun Wan. 2024b. Mc-mke: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. *arXiv preprint arXiv:2406.13219*.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.